FRONTIERS IN GENETICS
SPECIALTY GRAND CHALLENGE ARTICLE
published: 15 September 2011
doi: 10.3389/fgene.2011.00060

Grand challenges in bioinformatics and computational biology

Arcady Mushegian*

Stowers Institute Medical Research, Kansas City, MO, USA
**Correspondence: arm@stowers.org*

The terms "bioinformatics" and "computational biology" are only about 20 years old – their first appearance in the Medline database may have been as the keywords for the 1990 article describing the first steps of the National Center for Biotechnology Information (Benson et al., 1990). But the mathematical approaches toward the analysis of the structure, function, and evolution of biopolymers has begun much earlier. The field cannot be properly described without noting the series of articles by L. Pauling and E. Zuckerkandl in the 1960s, in which they cast evolutionary molecular biology as information science, called genes and their RNA and protein products "sense-carrying units" and defined the research program of distinguishing signals from noise in these molecules by comparing similar molecules from different species (Zuckerkandl and Pauling, 1965); Margaret Dayhoff's Atlases of protein sequence and structure and the first models of amino acid sequence evolution specified there (Dayhoff and Eck, 1968); first practical algorithms for pairwise comparisons of protein and nucleic acid sequence (Needleman and Wunsch, 1970; Sellers, 1974; Smith and Waterman, 1981); powerful database search heuristics, developed by David Lipman and his collaborators (Pearson and Lipman, 1988; Altschul et al., 1990); and the foundational work in algorithms and statistics for molecular phylogenetics (reviewed in Felsenstein, 2003). These efforts are (or should be) part of any textbook, not only on bioinformatics, but perhaps on modern biology as a whole.

The significance of this work, which was not yet called either bioinformatics or computational biology, was not only in the new discoveries of sequence similarities and patterns of gene evolution. A more profound effect was on the broader scientific enterprise, as it became obvious that computational methods were in fact instrumental in immediate and productive interpretation of the biological sequences.

Invention of automated sequencer in early 1980s, formal establishment of Human Genome Project in 1990, and announcement of the completion of human genome draft in 2000 are the recognized milestones of biology. The day of July 25, 1995, however, is a good candidate for the Day 1 of the Heroic Age of the Era of Complete Genomes. On that day, the first complete genome sequence of a cellular organism, *H. influenzae*, was reported in the *Science* magazine (Fleischmann et al., 1995). With several more genome sequences of bacteria and, later, archaea, the approaches of bioinformatics and computational biology could be applied at the genomic scale. And, although our understanding of the organisms with completely sequenced genomes continues to be far from perfect, a significant amount of new information could be teased out of the very first genomes in a matter of few months – the concern that the genome sequences for a long time will remain an enigma, similar to Linear A script of Cretan archeologists, proved to be unfounded.

Also in the 1980–1990s, the quiet revolutions occurred in many other fields of scientific instrumentation development, by now allowing us not only to sequence genomes completely, but also to profile quantitatively the amounts, activities, spatial locations, and movements of many molecules inside the cells, as well as register multiple parameters describing the whole cells and cell populations. Thus, in addition to the strings of symbols representing the genetic information encoded in the genomes, we have another genome-scale data type – the vectors of numeric measurements associated with every genetic element and every other molecule in the cells of different types, as well as with the cells and supracellular structures. In this case too, we are not completely clueless, as the existing algorithmic approaches and methods of multidimensional statistics help to discern biologically significant patterns in these data, and, on the other hand, the properties of these data motivate the development of new methods. It does not hurt that, as biologists come up with new platforms for data acquisition, the cost of high-performance computing and terabyte-scale data storage continues to go down.

To know any biological system, we want to get an insight in its evolution, structure, and function, in order to explain ultimately adaptation, diversity, and complexity of the system. Developing mathematical, computational, and statistical approaches and applying them to analyze these and other properties of living systems is the ultimate grand challenge for bioinformatics and computational biology.

More specific challenges run in several dimensions. First, there is a multiplicity of areas within biology, some of which already have the lists of open computational problems, and others so new that the problems are only now being defined. In particular, this journal is interested in the analysis of amino acid and nucleotide sequences, and in the novel views on the relationship between the sequences and the higher-order molecular structures; in the analysis of large multidimensional numeric datasets, which include gene expression readouts, gene and protein interaction data, and generally any representation in which each gene in a genome is associated with the set of measurements characterizing some aspect of the existence and life history of this gene; in computational analysis of evolution of all life forms; in quantitative approaches to analysis of biological images, biomedical texts, and other types of data that are only recently entering the real of high-throughput biology.

The second dimension of our quest is the dual existence of bioinformatics as enabling technology ("developing tools") and as a science that applies the tools of the trade to solving open problems in any of the areas described above. The journal will aim at providing the forum to the whole spectrum of studies, from an improvement of an existing method to an attempt at defining a new law (or perhaps at least a rule) of genomic biology. It will be important for us to see, however, that the description of a particular computational approach is

preceded by a clear definition of the scientific question to be answered using this approach, and that the success or failure of the new approach to answer the question and to move the research forward is clearly documented.

The third dimension of interest to our journal is epistemological. There is a sound argument, put forward many times, that the work carried out by using mostly or only dry-lab methods should be held to the same standards as the wet-lab work, and the conclusions from computational work are not inherently stronger or weaker than those from genetic or biochemical experiments. Wet-lab experiments are as amenable to alternative interpretations as the computational ones, and quantitative inference is inherent the interpretation of any experiment (Iyer et al., 2001). Thus, we will be interested in publishing the papers that deal with the standards of proof in computational biology, and in using wet-lab and dry-lab evidence to refute, but ultimately refine, each other.

Finally, bioinformatics and computational biology are relatively new fields, and we work to address the unsolved problems. From this may come a temptation to cast one's work as completely standalone effort, the result of out-of-the-blue burst of creativity. In contrast, attributed to the doyen of Russian historians Nikolai Karamzin is the conviction that the respect of one's forbearers is a citizen's virtue, and we may remember that the seminal PNAS contribution of Richard Bellman explicating the dynamic programming algorithm cited inspiration from the work of Kenneth Arrow and Abraham Wald (Bellman, 1952), and that T. Smith and M. Waterman in the paper about their eponymous algorithm credited Walter Goad for independently coming up with a very similar idea (Smith and Waterman, 1981). Following these examples, in this journal we will encourage authors to present the historic context of their contribution by enthusiastically accounting relevant results of their predecessors. Such treatment will be especially requested when the manuscripts are considered for elevation to the higher tiers in Frontiers Evaluation System.

REFERENCES

Altschul, S. F., Gish, W., Miller, W.,Myers, E. W., and Lipman, D. J. (1990). Basic local alignments search tool. *J. Mol. Biol.* 215, 403–410.

Bellman, R. (1952). On the theory of dynamic programming. *Proc. Natl. Acad. Sci. U.S.A.* 38, 716–719.

Benson, D., Boguski, M., Lipman, D., and Ostell, J. (1990). The national center for biotechnology information. *Genomics* 6, 389–391.

Dayhoff, M. O., and Eck, R. V. (1968). *Atlas of Protein Sequence and Structure*, Vol. 3. Silver Spring, MD: National Biochemical Research Foundation.

Felsenstein, J. (2003). *Inferring Phylogenies*. Sunderland, MA: Sinauer.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., Keith, M., Granger, S., Will, F., Chris, F., Jeannie, G. D., John, S., Robert, S., Li-Ing, L., Anna, G., Jenny, M. K., Janice, F. W., Cheryl, A. P., Tracy, S., Eva, H., Matthew, D. C., Teresa, R. U., Michael, C. H., David, T. N., Deborah, M. S., Rhonda, C. B., Leah, D. F., Janice, L. F., Joyce, L. F., Geoghagen, N. S. M., Cheryl, L. G., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., and Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512.

Iyer, L. M., Aravind, L., Bork, P., Hofmann, K., Mushegian, A. R., Zhulin, I. B., and Koonin, E. V. (2001). Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biol.* 2, Research0051.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.

Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 85, 2444–2448.

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* 26, 781–793.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.

Zuckerkandl, E., and Pauling, L. (1965). Molecules as documents of evolutionary history. *J. Theor. Biol.* 8, 357–366.

Received: 15 August 2011; accepted: 18 August 2011; published online: 15 September 2011.

Citation: Mushegian A (2011) Grand challenges in bioinformatics and computational biology. Front. Gene. ***2****:60. doi: 10.3389/fgene.2011.00060*

This article was submitted to Frontiers in Bioinformatics and Computational Biology, a specialty of Frontiers in Genetics.